What is Azure AI Content Safety?

Azure AI Content Safety is an AI service that detects harmful user-generated and AI-generated content in applications and services. Azure AI Content Safety includes text and image APIs that allow you to detect material that's harmful. The interactive Content Safety Studio allows you to view, explore, and try out sample code for detecting harmful content across different modalities.

Content filtering software can help your app comply with regulations or maintain the intended environment for your users.

Prerequisites

To use Azure AI Content Safety, you need:

- An Azure subscription - Create one for free
- A Content Safety resource created in a supported region

This documentation contains the following article types:

- Concepts provide in-depth explanations of the service functionality and features.
- Quickstarts are getting-started instructions to guide you through making requests to the service.
- How-to guides contain instructions for using the service in more specific or customized ways.

Content moderation features

Content Safety provides several APIs for different moderation needs:

AI safety and prompt protection

Expand table

Feature	Purpose	Concepts guide	Get started
Prompt Shields	Scans text for the risk of a User input attack on a Large Language Model.	Prompt Shields concepts	Quickstart
Groundedness detection (preview)	Detects whether the text responses of large language models (LLMs) are grounded in the source materials provided by the users.	Groundedness detection concepts	Quickstart
Protected material text detection	Scans AI-generated text for known text content (for example, song lyrics, articles, recipes, selected web content).	Protected material concepts	Quickstart
Task adherence API	Detects when tool use by AI agents is misaligned, unintended, or premature in the context of a user interaction.	Task adherence concepts	Quickstart

Content analysis

Expand table

Feature	Purpose	Concepts guide	Get started
Analyze text API	Scans text for sexual content, violence, hate, and self harm with multi-severity levels.	Harm categories	Quickstart
Analyze image API	Scans images for sexual content, violence, hate, and self harm with multi-severity levels.	Harm categories	Quickstart

Custom detection

Expand table

Feature	Purpose	Concepts guide	Get started
Custom categories (standard) API (preview)	Lets you create and train your own custom content categories and scan text for matches.	Custom categories concepts	Quickstart
Custom categories (rapid) API (preview)	Lets you define emerging harmful content patterns and scan text and images for matches.	Custom categories concepts	How-to guide

Where it's used

The following are a few scenarios in which a software developer or team would require a content moderation service:

- User prompts submitted to a generative AI service.
- Content produced by generative AI models.
- Online marketplaces that moderate product catalogs and other user-generated content.
- Gaming companies that moderate user-generated game artifacts and chat rooms.
- Social messaging platforms that moderate images and text added by their users.
- Enterprise media companies that implement centralized moderation for their content.
- K-12 education solution providers filtering out content that's inappropriate for students and educators.

> **Important**

You can't use Azure AI Content Safety to detect illegal child exploitation images.

Content Safety Studio

Azure AI Content Safety Studio is an online tool designed to handle potentially offensive, risky, or undesirable content using cutting-edge content moderation ML models. It provides templates and customized workflows, enabling users to choose and build their own content moderation system. Users can upload their own content or try it out with provided sample content.

Content Safety Studio not only contains out-of-the-box AI models but also includes **Microsoft's built-in terms blocklists** to flag profanities and stay up to date with new content trends. You can also upload your own blocklists to enhance the coverage of harmful content that's specific to your use case.

Studio also lets you set up a **moderation workflow**, where you can continuously monitor and improve content moderation performance. It can help you meet content requirements from all kinds of industries like gaming, media, education, E-commerce, and more. Businesses can easily connect their services to the Studio and have their content moderated in real-time, whether user-generated or AI-generated.

All of these capabilities are handled by the Studio and its backend; customers don't need to worry about model development. You can onboard your data for quick validation and monitor your KPIs accordingly, like technical metrics (latency, accuracy, recall), or business metrics (block rate, block volume, category proportions, language proportions, and more). With simple operations and configurations, customers can test different solutions quickly and find the best fit, instead of spending time experimenting with custom models or doing moderation manually.

Try Content Safety Studio

Content Safety Studio features

In Content Safety Studio, the following Azure AI Content Safety features are available:

- Moderate Text Content■: With the text moderation tool, you can easily run tests on text content. Whether you want to test a single sentence or an entire dataset, our tool offers a user-friendly interface that lets you assess the test results directly in the portal. You can experiment with different sensitivity levels to configure your content filters and blocklist management, ensuring that your content is always moderated to your exact specifications. Plus, with the ability to export the code, you can implement the tool directly in your application, streamlining your workflow and saving time.

- Moderate Image Content■: With the image moderation tool, you can easily run tests on images to ensure that they meet your content standards. Our user-friendly interface allows you to evaluate the test results directly in the portal, and you can experiment with different sensitivity levels to configure your content filters. Once you've customized your settings, you can easily export the code to implement the tool in your application.

- Monitor Online Activity■: The powerful monitoring page allows you to easily track your moderation API usage and trends across different modalities. With this feature, you can access detailed response information, including category and severity distribution, latency, error, and blocklist detection. This information provides you with a complete overview of your content moderation performance, enabling you to optimize your workflow and ensure that your content is always moderated to your exact specifications. With our user-friendly interface, you can quickly and easily navigate the monitoring page to access the information you need to make informed decisions about your content moderation strategy. You have the tools you need to stay on top of your content moderation performance and achieve your content goals.

Security

Microsoft Entra ID or Managed Identity

For enhanced security, you can use Microsoft Entra ID or Managed Identity (MI) to manage access to your resources.

- Managed Identity is automatically enabled when you create a Content Safety resource.
- Microsoft Entra ID is supported in both API and SDK scenarios. Refer to the general AI services guideline of Authenticating with Microsoft Entra ID. You can also grant access to other users within your organization by assigning them the roles of **Cognitive Services Users** and **Reader**. To learn more about granting user access to Azure resources using the Azure portal, refer to the Role-based access control guide.

Encryption of data at rest

Learn how Azure AI Content Safety handles the encryption and decryption of your data. Customer-managed keys (CMK), also known as Bring Your Own Key (BYOK), offer greater flexibility to create, rotate, disable, and revoke access controls. You can also audit the encryption keys used to protect your data.

Pricing

Azure AI Content Safety has an **F0** and **S0** pricing tier. See the Azure pricing page for more information.

Service limits

> **Important**
>
> **Deprecation Notice**
>
> As part of Content Safety versioning and lifecycle management, we're announcing the deprecation of certain Public Preview and GA versions of our service APIs. Following our deprecation policy:
>
> - **Public Preview versions**: Each new Public Preview version will trigger the deprecation of the previous preview version after a 90-day period, provided no breaking changes are introduced.
> - **GA versions**: When a new GA version is released, the prior GA version will be deprecated after a 90-day period if compatibility is maintained.
>
> See the What's new page for upcoming deprecations.

Input requirements

See the following list for the input requirements for each feature.

- **Analyze text API**:
  - Default maximum length: 10K characters (split longer texts as needed).
- **Analyze image API**:
  - Maximum image file size: 4 MB
  - Dimensions between 50 x 50 and 7200 x 7,200 pixels.
  - Images can be in JPEG, PNG, GIF, BMP, TIFF, or WEBP formats.
- **Analyze multimodal API (preview)**:
  - Default maximum text length: 1K characters.
  - Maximum image file size: 4 MB

    - Dimensions between 50 x 50 and 7200 x 7,200 pixels.
    - Images can be in JPEG, PNG, GIF, BMP, TIFF, or WEBP formats.
- **Prompt Shields API**:
    - Maximum prompt length: 10K characters.
    - Up to five documents with a total of 10K characters.
- **Groundedness detection API (preview)**:
    - Maximum length for grounding sources: 55,000 characters (per API call).
    - Maximum text and query length: 7,500 characters.
    - Minimum query length: 3 words.
- **Protected material detection APIs**:
    - Default maximum length: 10K characters.
    - Default minimum length: 110 characters (for scanning LLM completions, not user prompts).
- **Custom categories (standard) API (preview)**:
    - Maximum inference input length: 1K characters.
- **Task adherence (preview)**:
    - Maximum input length: 100K characters.

Language support

The Azure AI Content Safety models for protected material, groundedness detection, and custom categories (standard) work with English only.

Other Azure AI Content Safety models have been specifically trained and tested on the following languages: Chinese, English, French, German, Spanish, Italian, Japanese, Portuguese. However, these features can work in many other languages, but the quality might vary. In all cases, you should do your own testing to ensure that it works for your application.

For more information, see Language support.

Region availability

To use the Content Safety APIs, you must create your Azure AI Content Safety resource in a supported region. Currently, the Content Safety features are available in the following Azure regions with different API versions:

Expand table

Region	Custom Category (standard)	Groundedness	Image	Multimodal(Image with Tex)	Custom Category (rapid)	Prompt Shield	Protected Material (Text)	Protected Material (Code)	Text
Australia East	✅		✅		✅	✅	✅	✅	✅
Canada East			✅		✅	✅	✅	✅	✅
Central US			✅		✅	✅	✅	✅	✅
East US	✅	✅	✅	✅	✅	✅	✅	✅	✅
East US 2		✅	✅		✅	✅	✅	✅	✅
France Central		✅	✅		✅	✅	✅	✅	✅
Germany West Central					✅	✅	✅		✅
Italy North					✅	✅	✅		✅
Japan East			✅		✅	✅	✅	✅	✅
Korea Central			✅			✅	✅		✅
North Central US			✅		✅	✅	✅	✅	✅

Region	Custom Category (standard)	Groundedness	Image	Multimodal(Image with Tex)	Custom Category (rapid)	Prompt Shield	Protected Material (Text)	Protected Material (Code)	Text
Poland Central			✅			✅	✅	✅	✅
South Central US			✅		✅	✅	✅	✅	✅
South India			✅		✅	✅	✅	✅	✅
Sweden Central		✅	✅		✅	✅	✅	✅	✅
Switzerland North	✅		✅		✅	✅	✅	✅	✅
Switzerland West			✅		✅	✅	✅	✅	✅
UAE North			✅		✅	✅	✅	✅	✅
UK South		✅	✅		✅	✅	✅	✅	✅
West Europe			✅	✅	✅	✅	✅	✅	✅
West US		✅	✅		✅	✅	✅	✅	✅
West US 2			✅		✅	✅	✅	✅	✅
West US 3			✅		✅	✅	✅	✅	✅
FairFax - USGovArizona			✅			✅	✅	✅	✅
FairFax - USGovVirginia			✅			✅	✅		✅

Feel free to contact us if your business needs other regions to be available.

Query rates

Content Safety features have query rate limits in requests-per-second (RPS) or requests-per-10-seconds (RP10S). See the following table for the rate limits for each feature.

Expand table

Pricing tier	Moderation APIs (text and image)	Prompt Shields	Protected material detection	Groundedness detection (preview)	Custom categories (rapid) (preview)	Custom categories (standard) (preview)	Multimodal
F0	5 RPS	5 RPS	5 RPS	N/A	5 RPS	5 RPS	5 RPS
S0	1000 RP10S	1000 RP10S	1000 RP10S	50 RPS	1000 RP10S	5 RPS	10 RPS

If you need a faster rate, contact us to request it.

Contact us

If you get stuck, email us or use the feedback widget at the bottom of any Microsoft Learn page.

Next steps

Choose your development path:

Text moderation:

Text analysis quickstart

Image moderation:

Image analysis quickstart

Prompt protection:

Prompt Shields quickstart

ⓘ**Note:** The author created this article with assistance from AI. Learn more

---

Last updated on 01/31/2026